Protax-fungi : a web-based tool for probabilistic taxonomic placement of fungal internal transcribed spacer sequences

Abarenkov, Kessy

2018-10

Abarenkov , K , Somervuo , P , Nilsson , R H , Kirk , P M , Huotari , T , Abrego , N & Ovaskainen , O 2018 , ' Protax-fungi : a web-based tool for probabilistic taxonomic placement of fungal internal transcribed spacer sequences ' , New Phytologist , vol. 220 , no. 2 , pp. 517-525 . https://doi.org/10.1111/nph.15301

http://hdl.handle.net/10138/307605
https://doi.org/10.1111/nph.15301

publishedVersion

New Phytologist

Research

PROTAX-fungi: a web-based tool for probabilistic taxonomic placement of fungal internal transcribed spacer sequences

Kessy Abarenkov[1]*, Panu Somervuo[2]*, R. Henrik Nilsson[3,4], Paul M. Kirk[5], Tea Huotari[6], Nerea Abrego[6] and Otso Ovaskainen[2,7]

[1]Natural History Museum, University of Tartu, Vanemuise 46, Tartu 51014, Estonia; [2]Organismal and Evolutionary Biology Research Programme, University of Helsinki, PO Box 65, Helsinki FI-00014, Finland; [3]Department of Biological and Environmental Sciences, University of Gothenburg, Box 461, 405 30 Göteborg, Sweden; [4]Gothenburg Global Biodiversity Centre, Box 461, SE-405 30 Göteborg, Sweden; [5]Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3DS, UK; [6]Department of Agricultural Sciences, University of Helsinki, PO Box 27, Helsinki FI-00014, Finland; [7]Centre for Biodiversity Dynamics, Department of Biology, Norwegian University of Science and Technology, N-7491 Trondheim, Norway

Author for correspondence:
Kessy Abarenkov
Tel: +372 53 54 26 48
Email: kessy.abarenkov@ut.ee

Received: *23 April 2018*
Accepted: *30 May 2018*

New Phytologist (2018)
doi: 10.1111/nph.15301

Key words: annotation, data quality, environmental sequencing, fungi, identification tool, internal transcribed spacer (ITS), molecular species identification, probabilistic taxonomic assignment.

Summary

- Incompleteness of reference sequence databases and unresolved taxonomic relationships complicates taxonomic placement of fungal sequences. We developed PROTAX-fungi, a general tool for taxonomic placement of fungal internal transcribed spacer (ITS) sequences, and implemented it into the PLUTOF platform of the UNITE database for molecular identification of fungi.
- With empirical data on root- and wood-associated fungi, PROTAX-fungi reliably identified (with at least 90% identification probability) the majority of sequences to the order level but only around one-fifth of them to the species level, reflecting the current limited coverage of the databases.
- PROTAX-fungi outperformed the SINTAX and RDB classifiers in terms of increased accuracy and decreased calibration error when applied to data on mock communities representing species groups with poor sequence database coverage. We applied PROTAX-fungi to examine the internal consistencies of the Index Fungorum and UNITE databases. This revealed inconsistencies in the taxonomy database as well as mislabelling and sequence quality problems in the reference database. The according improvements were implemented in both databases.
- PROTAX-fungi provides a robust tool for performing statistically reliable identifications of fungi in spite of the incompleteness of extant reference sequence databases and unresolved taxonomic relationships.

Introduction

Fungi form a large and heterogeneous group of eukaryotic organisms of disparate ecological roles. Many fungi interact with plants through associations ranging from parasitic through saprotrophic to mutualistic. Advances in Next Generation DNA sequencing have made it possible to examine the nature of plant–fungal interactions in unprecedented detail (e.g. Eusemann *et al.*, 2016; Waring *et al.*, 2016). Even so, several shortcomings beset molecular identification of fungi and fungal communities. Importantly, only 2–6% of the estimated 2–6 million extant species of fungi have been described formally (Taylor *et al.*, 2014; Hawksworth & Lüking, 2017) and < 0.5% have been sequenced for the formal fungal barcode (Taylor *et al.*, 2014) – the nuclear ribosomal internal transcribed spacer (ITS) region. Furthermore, > 10% of the public fungal ITS sequences can be assumed to be misidentified (Nilsson *et al.*, 2012), further complicating molecular identification procedures. Although software solutions for molecular identification of fungi are available (e.g. RDP: Wang *et al.*, 2007; SINTAX: Edgar, 2016), they fail to account for the large number of fungal species for which we do not have reference data at present, and consequently they may not provide robust estimates of the reliability of the proposed taxonomic affiliations of the query sequences. These complications make it difficult to study fungal communities at the desired levels of resolution and scientific reproducibility.

In our previous work, we developed PROTAX (PRObabilistic TAXonomic placement; Somervuo *et al.*, 2016, 2017), a general statistical method for classifying DNA sequences according to specified taxonomy and reference databases. PROTAX yields statistically calibrated probabilities for taxonomic placement (Somervuo *et al.*, 2016), setting it apart from tools such as the RDP Classifier (Wang *et al.*, 2007) and SINTAX (Edgar, 2016), which rather offer heuristic estimates of the reliabilities of taxonomic assignments. PROTAX also accounts for the presence of undescribed (or described but unsequenced) lineages and mislabelled reference sequences. Here we introduce PROTAX-fungi,

*These authors contributed equally to this work.

which is a specific implementation of PROTAX aimed at classifying fungal ITS sequences. The purpose of PROTAX-fungi is to allow unbiased probabilistic assignment of fungal ITS sequences to known and unknown fungal lineages in a standardized and reproducible way. We specifically report on the implementation of PROTAX-fungi in the PLUTOF platform of the UNITE database for molecular identification of fungi (Kõljalg *et al.*, 2013; https://unite.ut.ee/). We examine the performance of the implementation on taxonomic placement of environmental fungal ITS sequences, as well use it to examine the internal consistency of the Index Fungorum+Species Fungorum and UNITE databases. Finally, we discuss the challenges involved in scientific examination of plant–fungal communities and provide a set of recommendations for the molecular ecology community.

Material and Methods

The Index Fungorum/Species Fungorum taxonomy database

PRObabilistic TAXonomic placement (PROTAX)-fungi relies on two reference datasets: a taxonomic classification system comprising all formally described species, and a reference sequence database whose entries typically only partially populate the taxonomy. The taxonomic classification system we use is Index Fungorum+Species Fungorum (http://www.indexfungorum.org), which exchanges data with MycoBank (Robert *et al.*, 2013) and Fungal Names (http://www.fungalinfo.net/), and is the most up-to-date resource for fungal names and classification. Using the current record names, we constructed a seven-level taxonomy. Incertae sedis classifications were replaced by creating dummy classifications along the path between the closest known lower and higher taxon nodes. The taxonomy database consists of 131 484 species classified into seven levels from kingdom to species (Table 1). For further information on Index Fungorum and Species Fungorum, see Supporting Information Methods S1.

The UNITE reference sequence database

As a reference sequence database, we used UNITE, an online database for reproducible molecular identification of fungi. All fungal sequences in UNITE are clustered into approximately species-level clusters which are called species hypotheses (SHs; Kõljalg *et al.*, 2013). All SHs are given a unique Digital Object Identifier (DOI; https://www.doi.org/) to promote unambiguous reference and communication across datasets and time, which would otherwise be hampered due to the frequent lack of precise Latin names for many fungal species. The use of DOIs also allows automatic assembly of metadata, such as host and country of collection (e.g. http://dx.doi.org/10.15156/BIO/SH181628.07FU). The SHs are subject to third-party annotation through the PLUTOF platform (Abarenkov *et al.*, 2010). The reference sequence database that we used consists of the 420 319 Sanger-derived sequences that comprised both the ITS1 and ITS2 sub-regions in v.7.1 of the SH system (https://unite.ut.ee/repository.php). Out of these reference sequences, 217 663 are annotated at the species level and they represent 17% of the described fungal species (Table 1).

The statistical model of PROTAX-fungi and its parameterization

PROTAX classifies query sequences in a hierarchical manner, starting at the root node of the taxonomy and proceeding towards the species level. The classification at each taxonomic node is conducted with a multinomial regression model that determines how the probability of 1.0 at the root node should be divided among the child nodes (Somervuo *et al.*, 2016, 2017). We constructed the predictors from pairwise sequence similarities between the query sequence and the reference sequences calculated by USEARCH (Edgar, 2010) and taxon membership given by SINTAX (Edgar, 2016) using USEARCH v.10.0.240_i86linux32. In total, the regression model has six predictors: (1) an indicator variable describing whether the node represents a known branch of the taxonomy or a taxon not present in Index Fungorum; (2) an indicator variable describing if any reference sequences are available; (3) the maximum similarity between the query sequence and the reference sequences; (4) the logarithm of the number of species nodes under the current node; (5) taxon membership obtained from the SINTAX classifier; and (6) the level of evidence that the best similarity might be due to mislabelling of the reference sequence. To define the last predictor, we note that cases where one reference sequence obtains a high similarity value and all

Table 1 Summary statistics of the taxonomy and reference sequences used by PROTAX-fungi

Taxonomy level	Number of annotated nodes	Number of dummy nodes	Nodes without refseqs (%)	Nodes with one refseq (%)	Nodes with two refseqs (%)	Nodes with at least three refseqs (%)	Number of refseqs
Kingdom	1	0	0	0	0	100	420 319
Phylum	8	2	40	0	0	60	380 553
Class	43	2062	87	4	2	8	373 737
Order	184	2668	81	4	2	13	367 562
Family	758	3344	70	6	3	21	350 676
Genus	10812	0	65	7	3	25	313 793
Species	131484	0	83	6	3	8	217 663

The number of reference sequences (refseqs) decreases from the kingdom level to the species level as not all reference sequences are annotated up to the species level. Dummy nodes were used to fill gaps (Incertae sedis) in taxonomy annotations. The number of nodes listed in the table does not include the additional branches that were added to each node to represent unknown taxa not included in Index Fungorum.

other reference sequences obtain a low similarity value suggest that the exceptionally well matching reference sequence may be mislabelled, especially if there are many poorly matching reference sequences. As a proxy for the level of evidence for the best match being due to mislabelling (predictor 6), we calculated the difference between the maximum similarity and the 97% quantile (excluding the maximum) similarity, and weighted the difference by $1-n^{-0.2}$, where n is the number of the reference sequences available for the node. In addition to the parameters describing the influences of the predictors 1–6, the model involves a parameter estimating the mislabelling probability of the training data (Somervuo *et al.*, 2016, 2017).

We estimated the model parameters separately for each level of the taxonomy, using the approach described in detail in Somervuo *et al.* (2016, 2017). Briefly, at each taxonomic level, we generated 5000 training sequences, of which 250 sequences (5%) represented missing taxonomic branches, resulting in 26% of missing species over the entire six-level taxonomy. The remaining training data represented species sampled randomly from the taxonomy, including cases with and without reference sequences. We trained separate models for three different input types: (1) the ITS1 alone, (2) the ITS2 alone, and (3) the full ITS region. Pairwise sequence similarity was computed using usearch_global (Edgar, 2010) with arguments –id 0.75 –maxaccepts 1000.

The implementation of PROTAX-fungi into PLUTOF

We implemented PROTAX-fungi in the PLUTOF platform. The PLUTOF platform is supported by the University of Tartu High Performance Computing Center (HPCC, http://www.hpc.ut.ee/), and it runs as a web-based sequence analysis environment where users can plan and carry out projects using a number of tools and resources. For conducting PROTAX-fungi analyses (Fig. 1), the user specifies input sequence data and parameters guiding the identification process (sequence region and probability threshold to be used for illustration purposes), after which (1) the new job is saved in the database, (2) the query sequences together with user-specified parameters are sent to HPCC, (3) the analysis is carried out, and (4) the results are returned to and stored in PLUTOF. Once the job is finished, an email notification with a link for download of the results is sent to the user. As standard output, PROTAX-fungi provides a Krona chart (Ondov *et al.*, 2011) that allows the user to explore the classifications and their reliabilities through an interactive graphical interface (Fig. 1c), and the same information in numerical format, including for each query sequence both the probabilistic classifications and information on the two best matches to the reference sequence database and on their corresponding SHs when available. The source code of PROTAX-fungi is deposited in Github (https://github.com/psomervuo/protaxfungi).

Building mock communities

We tested the performance of PROTAX-fungi and compared it to SINTAX and RDP classifiers by applying them to artificially generated mock communities. To ensure the comparability of the results, we used identical input data (reference sequence and taxonomy databases) for all three classifiers. With the mock communities, we attempted to mimic environmental sequence data. Such data can be expected to vary greatly in how well the taxonomy and reference databases cover them, depending on, for example, the geographical location and substrate sampled. To consider a range of situations, we included six scenarios (Table 2), which relate to the extent to which the species behind the environmental sequences are included in the taxonomy database (well-known or poorly known species groups), and how well they are represented by reference sequences (species groups with high, intermediate, or poor coverage of reference sequences).

Index Fungorum comprises 131 484 species, which is only a small fraction (2–6%) of the estimated number of 2–6 million fungal species (Taylor *et al.*, 2014; Hawksworth & Lüking, 2017). However, the species that have not been described to science may be on average less common than those that have been described already, so it is plausible that a larger proportion of species in an environmental sample are known to science, especially if the sample originates from a much studied substrate type or if the research focuses on a well-known taxonomical group. Thus, when generating a mock community representing a well-known species group (say, wood-inhabiting fungi from boreal forests), we assumed that 90% of the species included in the mock community are present in the taxonomy database. When generating mock communities representing a poorly known species groups (say, soil fungi from tropical forests), we assumed that 25% of the species included in the mock community are present in the taxonomy database.

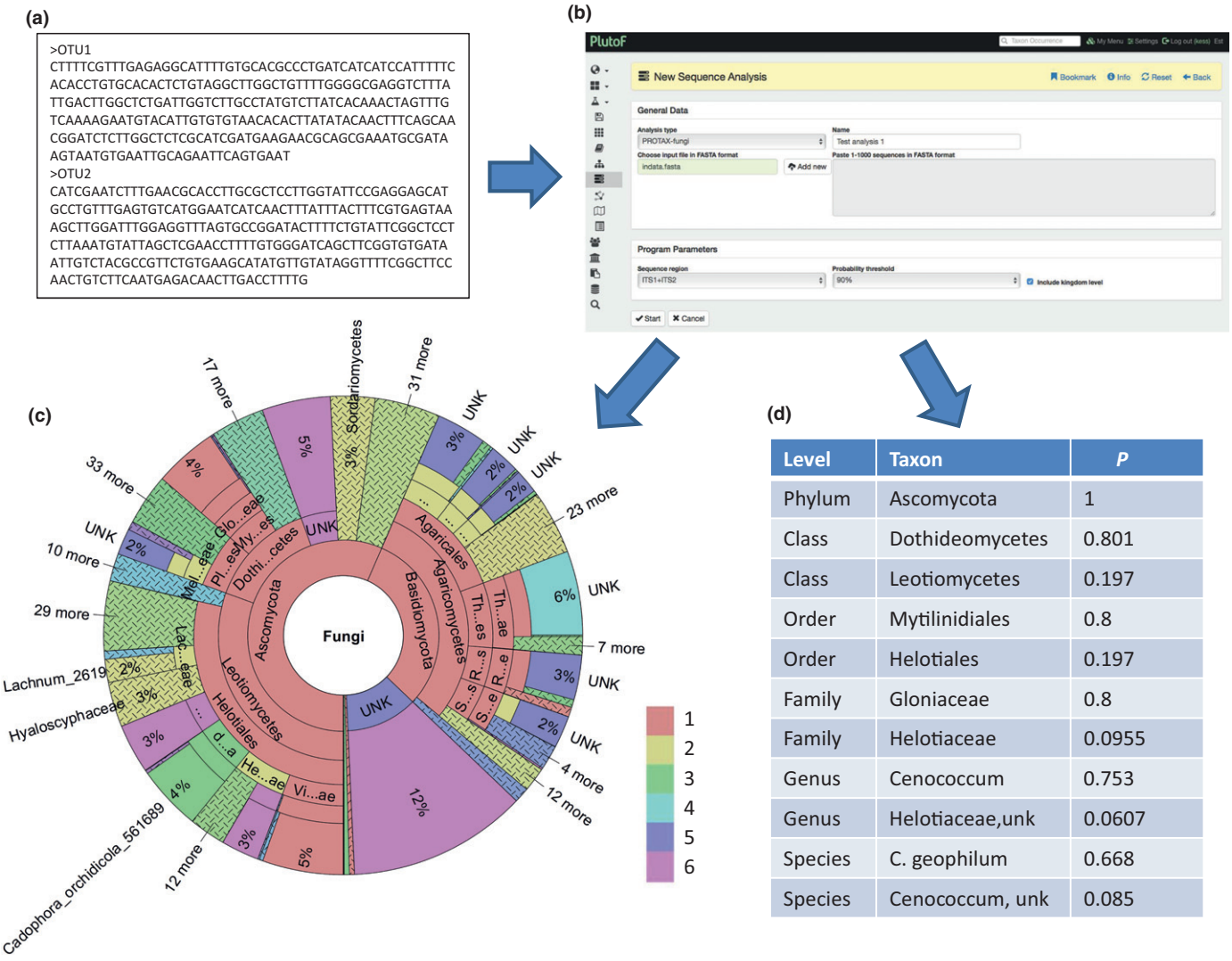
The reference sequence database includes at least one sequence for 17% of those species that are included in the database. However, the coverage of reference sequences may be higher for common species than for rare species. When generating mock communities representing species groups with high (respectively, intermediate or poor) coverage of reference sequences, we assumed that 90% (respectively, 50% or 20%) of the species had at least one reference sequence.

We generated the mock communities artificially by extracting sequences from the reference sequence database (as in Halwachs *et al.* (2017) and Motooka *et al.* (2017)) rather than, for example, sequencing species mixtures (as in Bjørnsgaard *et al.* (2017) or Bakker (2018)). The rationale is that we wanted to exercise control over the structures of the communities in terms of their proportions of species unknown to science or species without reference sequences, as we assume both sources of uncertainty to be highly relevant for most eDNA datasets (see earlier). All six mock communities were constructed to comprise 1000 species, and we constructed 10 replicates for each of the cases (Table 2). Technical details on how the mock communities were built, including removal of taxonomical branches to create unknown taxa, are provided in Methods S2.

Comparing the performance of PROTAX-fungi with SINTAX and RDP

We applied the PROTAX-fungi, SINTAX and RDP classifiers to both ITS1 and ITS2 data on the mock communities, and

New Phytologist



Level	Taxon	*P*
Phylum	Ascomycota	1
Class	Dothideomycetes	0.801
Class	Leotiomycetes	0.197
Order	Mytilinidiales	0.8
Order	Helotiales	0.197
Family	Gloniaceae	0.8
Family	Helotiaceae	0.0955
Genus	Cenococcum	0.753
Genus	Helotiaceae,unk	0.0607
Species	C. geophilum	0.668
Species	Cenococcum, unk	0.085

Fig. 1 An overview of the PROTAX-fungi pipeline implementation in the PLUTOF platform. The user uploads query sequences (a) and sets parameters guiding the identification (b). The output consists of probabilistic classifications of the query sequences, and is provided both in an interactive Krona chart (c) and in numerical format (d). Panels (c) and (d) are based on root-associated fungal data from eastern Greenland, with (c) showing all data and (d) probabilistic classification of a single sequence (only cases with probability > 0.05 are shown). In (c), the width of each sector is proportional to the expected number of sequences that was assigned to that taxonomic unit. The colours show the type and confidence level of each identification. Colours 1–3 correspond to well-identified taxonomic units for which the proportion of reliable identifications is in the range [50%...100%] (Colour 1), (0%...50%) (Colour 2), or 0% (Colour 3). Colours 4–6 correspond to unknown taxonomic units for which the proportion of reliable identifications is in the range [50%...100%] (Colour 4), (0%...50%) (Colour 5), or 0% (Colour 6). The same graph is provided as an interactive web page in https://github.com/psomervuo/protaxfungi.

compared their suggested taxonomic placements to the true labels to evaluate their performance. Although SINTAX and RDP usually return only the most likely placement for each taxonomical level, PROTAX-fungi returns classification probabilities for multiple taxonomic placements. To ensure an easy comparison among the methods, we utilized only the most likely classification also in the case of PROTAX-fungi. We characterized the quality of the taxonomic placements by measuring their accuracy and calibration error for each taxonomic level. With accuracy we intended to measure the frequency of correct taxonomic placements: how often the most likely classification is actually correct. With calibration error we intended to measure the validity of the classification probabilities. For example, if 100 sequences are assigned a classification probability of 0.8, the classification probabilities are well-calibrated if 80 of them are correct and the remaining 20 incorrect. We calculated accuracy as the number of correctly classified sequences divided by the total number of classified sequences. Cases for which correct classification was not defined (e.g. species-level classification for a sequence representing a missing genus) were excluded from the assessment of accuracy. If the classifier provided no classification, we considered the classification to be incorrect. The RDP classifier requires that all reference sequences have annotations for all levels in the taxonomy. If a reference sequence did not contain e.g. species-level annotation, we created such an annotation using a dummy name, and considered classifications

Table 2 Mock communities used for comparing the performances of the PROTAX-fungi, RDP and SINTAX classifiers

Mock community	Number of species	Reference sequence coverage	Taxonomical coverage
A1	1000	High (90%)	Well-known taxa (90% of the species included in taxonomy)
A2	1000	Intermediate (50%)	
A3	1000	Poor (20%)	
B1	1000	High (90%)	Poorly known taxa (25% of the species included in taxonomy)
B2	1000	Intermediate (50%)	
B3	1000	Poor (20%)	

Each mock community was constructed by selecting 1000 reference sequences each representing a different species, and then subsampling fungal taxonomy and the reference sequence database so that some of the mock species represented unknown taxa or taxa without reference sequences (see Supporting Information Methods S2 for details). In cases with high, intermediate or poor references sequence coverage, at least one reference sequence is available (respectively) for 90%, 50% or 20% of those mock species that are included in the taxonomy.

to dummy names as incorrect. The logic here is that if the best match is obtained to a sequence that is not annotated to species level, for example, the user has no information on what is the species behind the query sequence, and thus the classification is not accurate at that level. To evaluate the calibration error of the classification probabilities, we used all sequences for which the classifier produced results. The results were sorted based on the classification probabilities and divided into 10 equally sized bins. We calculated the mean absolute difference between the sum of classification probabilities and the number of correct classifications for each bin and then averaged the results.

Using PROTAX-fungi to classify environmental sequence data

We applied PROTAX-fungi to environmental ITS2 data originating from two datasets: sawdust samples from 100 spruce logs in a boreal forest in Finland (Ovaskainen *et al.*, 2013, henceforth called data on wood-associated fungi), and plant-root samples along an altitudinal gradient in eastern Greenland (called henceforth data on root-associated fungi). The Greenland data have been deposited to the Dryad data repository (doi: 10.5061/dryad.9dr6j0c) which is described in Methods S3. To reduce computational load, we clustered the sequences to operational taxonomic units (OTUs) with 98.5% clustering threshold, after which the most common sequence in each cluster was selected as the query sequence for PROTAX-fungi. In the post-processing phase, OTUs that received similar classification by PROTAX-fungi were merged.

We computed for both datasets the fraction of sequences that could be classified reliably (with at least 90% probability) or plausibly (with at least 50% probability) to a given taxonomical level, as well as the number of distinct taxonomic units that could be reliably or plausibly identified for each taxonomical level. We expected that the wood-associated fungi from Finland would be better represented in the taxonomy and reference sequence databases than the root-associated fungi from Greenland, and thus that the classification accuracy would be higher for data on wood-associated fungi.

Using PROTAX-fungi to reveal consistency problems with respect to the Index Fungorum+Species Fungorum and UNITE databases

We applied PROTAX-fungi to perform a taxonomic placement of all UNITE reference sequences, and compared these taxonomic placements to the label of the reference sequence. Following the test that Somervuo *et al.* (2016) performed with simulated data, we considered a reference sequence potentially inconsistent if the model predicted an incorrect taxonomic affiliation with high confidence. To be conservative, we required the probability of the most likely classification to be a least 100 times the probability of the outcome corresponding to the taxonomic affiliation of the sequence. This test was applied at all levels of the taxonomy. For all suspicious cases identified at higher taxonomic levels (phylum, class and order) and 50 randomly selected cases at lower levels (family, genus and species), we used expert evaluation to assess whether there was an actual problem or not, and tried to track the likely origin of the problem. We classified each potential inconsistency pointed out by PROTAX-fungi as: (i) no evidence of problem, if there was not clear evidence of a problem; (ii) 'mislabelling of a UNITE reference sequence'; (iii) 'sequence quality related issue'; (iv) 'taxonomy related issue'; or (v) as 'unclassified problem', where in the last category the expert was not sure where the actual problem was.

Results

Figure 1 illustrates the input and output of PROTAX-fungi, using the root-associated fungal data as an example. The classification results are graphically illustrated in the form of a Krona chart (Fig. 1c), showing a high coverage of the classes Leotiomycetes, Dothideomycetes and Agaricomycetes. Although defining a probability threshold (such as 90%) is necessary for colouring the Krona wheel according to the confidence level of the identifications, the numerical output of PROTAX-fungi is based on exact probability values and thus does not require the user to define any *a priori* threshold value. For example, in the probabilistic classification of a single OTU illustrated in Fig. 1(d), the most likely species *Cenococcum geophilum* has been assigned the probability of 0.668, a value that can be used, for example, in downstream analyses as the weight of how much this identification can be trusted. In this case, there is substantial uncertainty already at the class level, as the sequence received non-negligible probabilities for both of the classes Dothideomycetes and Leotiomycetes (Fig. 1d). Concerning computational efficiency, in a test run it took 10 min from the user submitting 500 sequences to receiving the email with results, out of which 4 min went to the actual processing of the sequences. The exact processing times may, however, vary depending on the present computational load at the HPCC.

The comparison between PROTAX-fungi and the RPD and SINTAX classifiers shows only minor differences at the phylum, class, order and family levels (Notes S1), moderate differences at the genus level (Fig. 2a,b), and major differences at the species level (Fig. 2c,d). The largest differences between the classifiers

appear with mock communities that represent poorly known species communities (mock communities B1, B2 and B3) and have a poor sequence coverage in the reference database (mock communities A3 and B3). For the latter cases, PROTAX-fungi has clearly higher accuracy and lower calibration error than RDP or SINTAX, especially at the species level. The comparison provided similar results for ITS1 and ITS2 sequences (Notes S1).

The application of PROTAX-fungi to the two empirical datasets shows how the proportion of sequences that can be reliably or plausible classified decreases with increasing taxonomic resolution. For example, although the majority of sequences could be reliably classified at the order level, only some 20% of them could be reliably classified to the species level (Fig. 3a). For the root-associated fungal data the proportion of sequences that cannot be classified even plausibly is larger than for the wood-associated fungal data (Fig. 3a), suggesting a higher proportion of taxa that are poorly covered in the reference databases for the root-associated data. In terms of diversity, as measured by the number of well-identified taxonomical units, the two dataset behave very similarly except at the species level (Fig. 3b). At this level, many more taxa could be identified for the wood-associated fungal dataset (Fig. 3b), again suggesting the presence of poorly covered taxa in the root-associated fungal data.

We used PROTAX-fungi to examine the consistency of 210 064 reference sequences included in UNITE. Out of these, 15 293 (7-%) indicated strong evidence for mislabelling. The majority of these cases (83%) occurred at species level (Fig. 4a). Manual identification of potentially mislabelled sequences verified the presence of a problem for most cases at the phylum, class, order and family levels, whereas a large proportion of ambiguous cases (no evidence of problem category, in Fig. 4b) remained at the genus and species levels. Especially at the levels of order and family, the most common problem was related to the taxonomic classification (Fig. 4b), such as misclassification of genera or higher taxa, or incorrect synonymy. For examples of misclassifications and their causes, see Notes S2.

Discussion

In the present paper we outline an approach for robust, probabilistic taxonomic assignment of internal transcribed spacer (ITS) sequences derived from plant-associated or other fungal communities. Our approach combines information from the Index Fungorum/Species Fungorum and the UNITE databases, and it accounts for the presence of unknown as well as known but unsequenced fungal lineages. Most importantly, PRObabilistic TAXonomic placement (PROTAX)-fungi yields as output not only the reference sequence that matches the best with the query sequence, but also the entire set of possible taxonomical affiliations and their probabilities. Thus, the user will know not only the most likely taxonomical affiliations, but also how much they can be trusted. We hope that this will serve as a safeguard against the all-too-common over-optimistic taxonomic assignments often done based on, for example, BLAST searches. As shown by our examples with root- and wood-associated fungal data, it is often difficult to obtain reliable species- or even genus-level assignments, largely due to the incompleteness of the current reference databases. In such a case, a robust order-level assignment is clearly to be preferred over spurious inference at the species or genus levels.

As revealed by the performance comparison with mock communities, PROTAX-fungi performs similarly to the RDP and SINTAX classifiers for datasets that are well covered by the taxonomy and reference sequence databases, but it performs better than the RDP and SINTAX classifiers for datasets that are poorly covered by the taxonomy and reference sequence databases. This is because PROTAX-fungi explicitly models the possibility of the test sequence belonging to an unknown taxonomical unit, or a

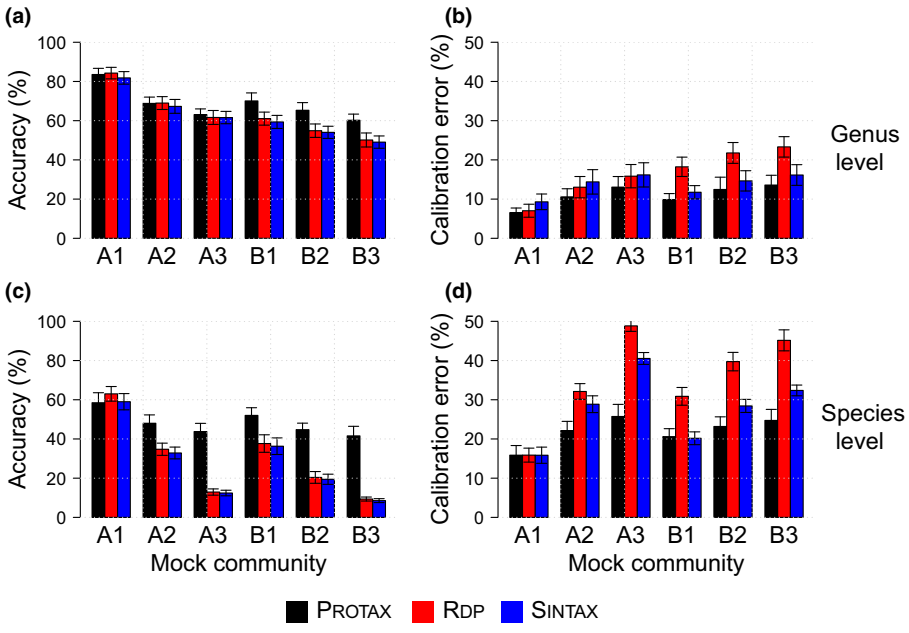


Fig. 2 Comparison among the performances of the PROTAX-fungi, RDP and SINTAX classifiers. The panels show the accuracy (a, c) and calibration error (b, d) for the three classifiers, as measured at the genus (a, b) and species (c, d) levels. The mock communities A1–A3 and B1–B3 are described in Table 2. The bars show the mean result and the error bars ± 1 SE over the 10 replicates for each of the mock communities. This figure shows results for the classifications based on the internal transcribed spacer (ITS)2. Corresponding results for the ITS1 region, as well classification results at the levels of phylum, class, order and family, are shown in Supporting Information Figs S1 and S2.

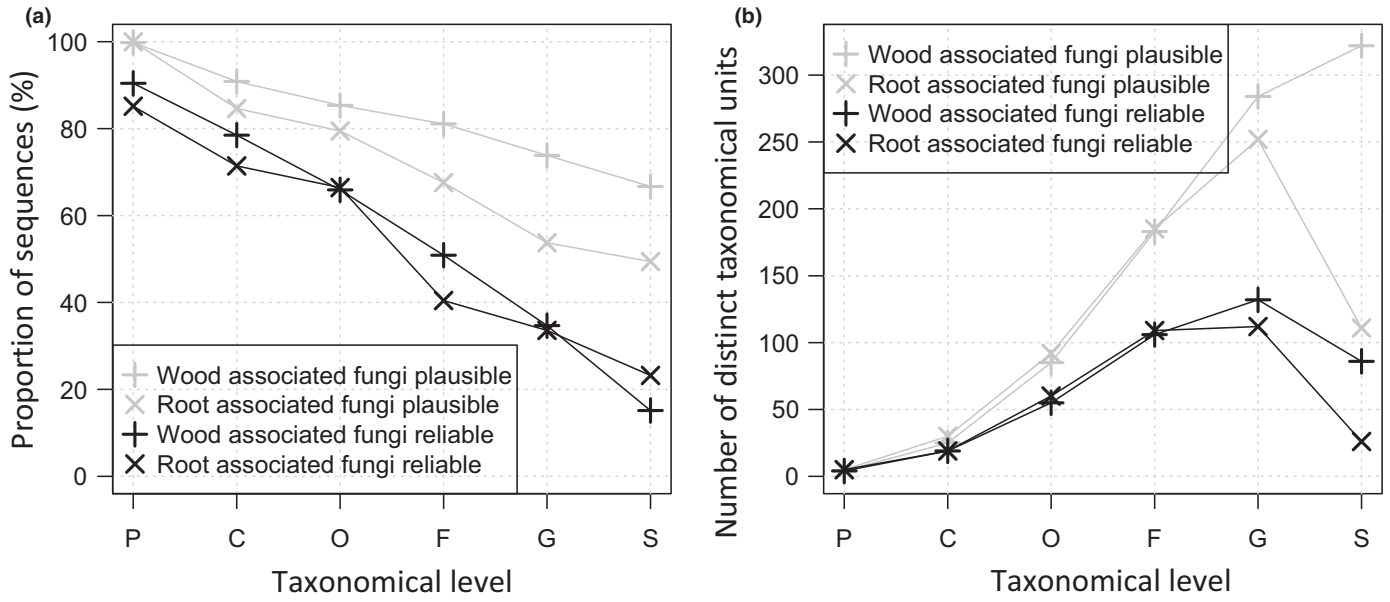


Fig. 3 The performance of Protax-fungi in classifying environmental sequences for the root- and wood-associated fungal datasets. (a) The proportion of sequences that could be classified either reliably (with at least 90% classification probability) or plausibly (with at least 50% classification probability) at each taxonomical level. (b) The number of distinct well-identified taxonomical units that could be identified either reliably or plausibly.

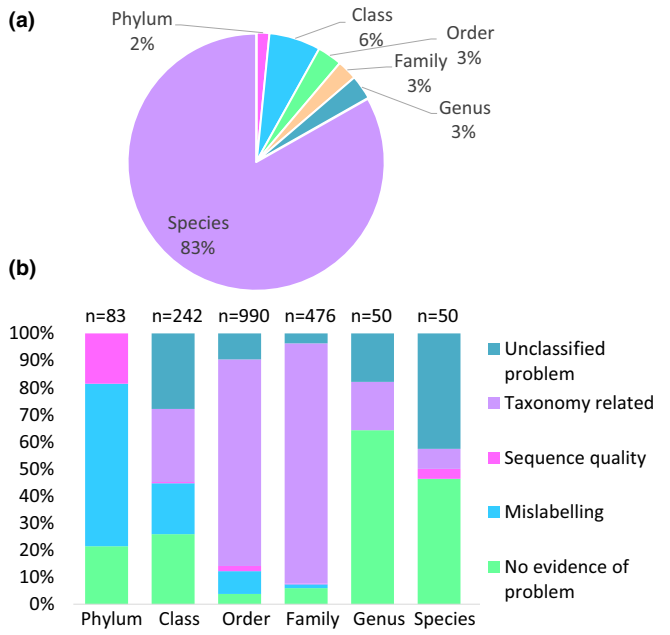


Fig. 4 Potentially compromised entries in UNITE identified by Protax-fungi. (a) The proportions of UNITE reference sequences for which Protax-fungi indicated strong evidence of mislabelling at a particular taxonomic level. (b) The result of an expert evaluation (by K. Abarenkov) examining the most likely reason why Protax-fungi identified a sequence problematic. The colours show the proportion of sequences that the expert considered to belong to each category, and the numbers on top of the bars show the number of sequences evaluated manually. Note that at family, genus and species levels only a subsample of 50 sequences was evaluated.

taxonomical unit without reference sequences, whereas the Rdp and Sintax classifiers either yield the taxonomical unit with the best-matching reference sequences, or fail to give any classification. Yet, we note that Protax-fungi should not be considered as a competitor to other classifiers, but rather as a statistically calibrated method that can utilize other classifiers as its predictors. As a case in point, we have used here the classification output by Sintax as one of six predictors in Protax-fungi. This makes it possible to utilize the high classification accuracy of Sintax for cases with high reference sequence coverage, but at the same time (through the other predictors) robustly account for the high amount of uncertainty generated by missing taxa or missing reference sequences.

Protax-fungi performs probabilistic identifications based on evaluating the query sequences against information that can be derived from the Index Fungorum+Species Fungorum and UNITE databases. However, it currently ignores any other kinds of information that may be available for refining the identification probabilities, leaving room for improving the classifications in the post-processing step. To illustrate this, assume that Protax-fungi assigns the probability 0.7 for taxon A, the probability 0.2 for taxon B and the probability 0.1 for all other possibilities. Assume further that, based on information on the geographical locality of sampling, the habitat, the substrate type, any interacting taxa, or other such information, the researcher can exclude the possibility of taxon B. Conditional on the identification not being taxon B, the user can then refine the probabilities to $0.7/0.8 = 0.875$ for taxon A and $0.1/0.8 = 0.125$ for all other possibilities, thus reducing the uncertainty. Given increasing availability of databases for plant and fungal traits (Kattge *et al.*, 2011; Treseder & Lennon, 2015; Nguyen *et al.*, 2016), there is increasing potential for utilizing such information in the identification of fungal communities from ecological studies.

In addition to taxonomic classification of environmental data, Protax-fungi can be used as a systematic and automated tool for identifying consistency problems in reference databases. Being able

to systematically identify and correct spurious entries from the databases is important, because incorrect ecological or functional assignments resulting from compromised species names may propagate through the literature and have considerable negative downstream repercussions (e.g. Gilks *et al.*, 2002; Kang *et al.*, 2010). Any shortcomings in UNITE and Index Fungorum+Species Fungorum are open for correction and to some extent third-party annotation, and we hope that PROTAX-fungi will provide a helpful tool for users to identify potentially problematic cases.

Molecular identification of fungi will remain difficult in the short term. Sequencing efforts in soil and freshwater systems, as well as in built environments, have highlighted a large range of new, undescribed fungi (e.g. Tedersoo *et al.*, 2014; Grossart *et al.*, 2016; Nilsson *et al.*, 2016). Few of these lineages are associated with known fruiting bodies or other somatic structures, and many of them cannot be kept in culture. This currently precludes formal description of most of these species and lineages (Hawksworth *et al.*, 2016; Nilsson *et al.*, 2016; Hibbett *et al.*, 2017; but see Rosling *et al.*, 2011), and thus the mycological community will simply have to live with the fact that many species known from sequence data will not have formal Latin names for some time to come. The UNITE SH system offers a means for unambiguous communication of these lineages, and PROTAX-fungi assists researchers in avoiding over-classification of newly generated fungal sequences into known, but incorrect species names. In this study we have contributed a software platform to help the molecular ecology community to apply PROTAX-fungi to analyse their sequence data at the level where robust conclusions can be drawn.

Acknowledgements

The research was funded by the Academy of Finland (grants 284601, 1273253 and 250444 to O.O., grant 285803 to T.H. and grant 308651 to N.A.) and the Research Council of Norway (CoE grant 223257). K.A. acknowledges support from the Estonian Research Council (IUT20-30) and the European Regional Development Fund (Centre of Excellence EcolChange). K.A., R.H.N. and the UNITE community acknowledge support from the Alfred P. Sloan Foundation.

Author contributions

K.A., P.S. and P.M.K implemented the software and performed data analysis; O.O. and R.H.N. assisted with additional data analysis and interpretation; N.A. and O.O. collected the empirical data and T.H. performed the laboratory work related to the case study on root-associated fungi; K.A., O.O., N.A., P.S., R.H.N. and P.M.K. wrote the manuscript; and all co-authors contributed to the revisions of the manuscript. K.A. and P.S. contributed equally to this work.

References

Abarenkov K, Tedersoo L, Nilsson RH, Vellak K, Saar I, Veldre V, Parmasto E, Prous M, Aan A, Ots M *et al.* 2010. PlutoF – a web-based workbench for ecological and taxonomical research, with an online implementation for fungal ITS sequences. *Evolutionary Bioinformatics* **6**: 189–196.

Bakker MG. 2018. A fungal mock community control for amplicon sequencing experiments. *Molecular Ecology Resources* **18**: 541–556.

Bjørnsgaard AA, Davey ML, Kauserud H. 2017. ITS all right mama: investigating the formation of chimeric sequences in the ITS2 region by DNA metabarcoding analyses of fungal mock communities of different complexities. *Molecular Ecology Resources* **17**: 730–741.

Edgar R. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–2461.

Edgar R. 2016. SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *bioRxiv* doi: 10.1101/074161.

Eusemann P, Schnittler M, Nilsson RH, Jumpponen A, Dahl MB, Würth DG, Buras A, Wilmking M, Unterseher M. 2016. Habitat conditions and phenological tree traits overrule the influence of tree genotype in the needle mycobiome–*Picea glauca* system at an arctic treeline ecotone. *New Phytologist* **211**: 1221–1231.

Gilks WR, Audit B, De Angelis D, Tsokas S, Ouzounis CA. 2002. Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics* **18**: 1641–1649.

Grossart HP, Wurzbacher C, James TY, Kagami M. 2016. Discovery of dark matter fungi in aquatic ecosystems demands a reappraisal of the phylogeny and ecology of zoosporic fungi. *Fungal Ecology* **19**: 28–38.

Halwachs B, Madhusudhan N, Krause R, Nilsson RH, Moissl-Eichinger C, Högenauer C, Thallinger GG, Gorkiewicz G. 2017. Critical issues in mycobiota analysis. *Frontiers in Microbiology* **8**: 180.

Hawksworth DL, Hibbett DS, Kirk PM, Lücking R. 2016. Proposals to permit DNA sequence data to serve as types of names of fungi. *Taxon* **65**: 899–900.

Hawksworth DL, Lüking R. 2017. Fungal diversity revisited: 2.2 to 3.8 million species. *Microbiology Spectrum* **5**: FUNK-0052-2016.

Hibbett D, Abarenkov K, Kõljalg U, Öpik M, Chai B, Cole J, Wang Q, Crous P, Robert V, Helgason T *et al.* 2017. Sequence-based classification and identification of Fungi. *Mycologia* **108**: 1049–1068.

Kang S, Mansfield MA, Park B, Geiser DM, Ivors KL, Coffey MD, Grünwald NJ, Martin FN, Lévesque CA, Blair JE. 2010. The promise and pitfalls of sequence-based identification of plant pathogenic fungi and oomycetes. *Phytopathology* **100**: 732–737.

Kattge J, Diaz S, Lavorel S, Prentice IC, Leadley P, Bönisch G, Garnier M, Westoby M, Reich PB, Wright IJ *et al.* 2011. TRY – a global database of plant traits. *Global Change Biology* **17**: 2905–2935.

Kõljalg U, Nilsson RH, Abarenkov K, Tedersoo L, Taylor AF, Braham M, Bates ST, Bruns TD, Bengtsson-Palme J, Callaghan TM *et al.* 2013. Towards a unified paradigm for sequence-based identification of Fungi. *Molecular Ecology* **22**: 5271–5277.

Motooka D, Fujimoto K, Tanaka R, Yaguchi T, Gotoh K, Maeda Y, Furuta Y, Kurakawa T, Goto N, Yasunaga T *et al.* 2017. Fungal ITS1 deep-sequencing strategies to reconstruct the composition of a 26-species community and evaluation of the gut mycobiota of healthy Japanese individuals. *Frontiers in Microbiology* **8**: 238.

Nguyen NH, Song Z, Bates ST, Branco S, Tedersoo L, Menke J, Schilling JS, Keppedy PG. 2016. FUNGuild: an open annotation tool for parsing fungal community datasets by ecological guild. *Fungal Ecology* **20**: 241–248.

Nilsson RH, Tedersoo L, Abarenkov K, Ryberg M, Kristiansson E, Hartman M, Schoch CL, Nylander JAA, Bergsten J, Porter TM *et al.* 2012. Five simple guidelines for establishing basic authenticity and reliability of newly generated fungal ITS sequences. *MycoKeys* **4**: 37–63.

Nilsson RH, Wurzbacher C, Bahram M, Coimbra VRM, Larsson E, Tedersoo L, Eriksson J, Duarte C, Svantesson S, Sánchez-Garciá M *et al.* 2016. Top 50 most wanted fungi. *MycoKeys* **12**: 29–40.

Ondov B, Bergman N, Phillippy A. 2011. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* **12**: 385.

Ovaskainen O, Schigel D, Ali-Kovero H, Auvinen P, Paulin L, Nordén B, Nordén J. 2013. Combining high-throughput sequencing with fruit-body surveys reveals contrasting life-history strategies in fungi. *ISME Journal* **7**: 1696–1709.

Robert V, Vu D, Amor AB, van de Wiele N, Brouwer C, Jabas B, Szoke S, Dridi A, Triki M, Bend Daoud S *et al.* 2013. MycoBank gearing up for new horizons. *IMA Fungus* **4**: 371–379.

Rosling A, Cox F, Cruz-Martinez K, Ihrmark K, Grelet GA, Lindahl BD, Menkis A, James TY. 2011. Archaeorhizomycetes: unearthing an ancient class of ubiquitous soil fungi. *Science* **333**(6044): 876–879.

Somervuo P, Koskela S, Pennanen J, Nilsson RH, Ovaskainen O. 2016. Unbiased probabilistic taxonomic classification for DNA barcoding. *Bioinformatics* **32**: 2920–2927.

Somervuo P, Yu D, Xu C, Ji Y, Hultman J, Wirta H, Ovaskainen O. 2017. Quantifying uncertainty of taxonomic placement in DNA barcoding and metabarcoding. *Methods in Ecology and Evolution* **8**: 398–407.

Taylor DL, Hollingsworth TN, McFarland JW, Lennon NJ, Nusbaum C, Ruess RW. 2014. A first comprehensive census of fungi in soil reveals both hyperdiversity and fine-scale niche partitioning. *Ecological Monographs* **84**: 3–20.

Tedersoo L, Bahram M, Põlme S, Kõljag U, Yorou NS, Wijesundera R, Ruiz LV, Vasco-Palacios AM, Thu PQ, Suija A *et al.* 2014. Global diversity and geography of soil fungi. *Science* **346**: 1078.

Treseder KK, Lennon JT. 2015. Fungal traits that drive ecosystem dynamics on land. *Microbiology and Molecular Biology Reviews* **79**: 243–262.

Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the New Bacterial Taxonomy. *Applied and Environmental Microbiology* **73**: 5261–5267.

Waring BG, Adams R, Branco S, Powers JS. 2016. Scale-dependent variation in nitrogen cycling and soil fungal communities along gradients of forest composition and age in regenerating tropical dry forests. *New Phytologist* **209**: 845–854.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article:

Fig. S1 Comparison among the performances of the PROTAX-fungi, RDP and SINTAX classifiers based on the ITS2 region.

Fig. S2 Comparison among the performances of the PROTAX-fungi, RDP and SINTAX classifiers based on the ITS1 region.

Methods S1 Information on the Index Fungorum + Species Fungorum databases

Methods S2 Technical details on how the mock communities were built

Methods S3 Description of the Greenland data on root-associated fungi

Notes S1 Comparing the performance of PROTAX-fungi with SINTAX and RDP

Notes S2 Examples of consistency problems with respect to the Index Fungorum+Species Fungorum and UNITE databases

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.